BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59228
doi: 10.3897/biss.4.59228

OPEN ACCESS

Conference Abstract

A Google Sheet Add-on for Biodiversity Data Standardization and Sharing

José Augusto Salim‡, Antonio Mauro Saraiva‡

‡ Universidade de São Paulo, São Paulo, Brazil

Corresponding author: José Augusto Salim (joseasalim@usp.br)

Received: 01 Oct 2020 | Published: 02 Oct 2020

Citation: Salim JA, Saraiva AM (2020) A Google Sheet Add-on for Biodiversity Data Standardization and Sharing. Biodiversity Information Science and Standards 4: e59228. https://doi.org/10.3897/biss.4.59228

Abstract

For those biologists and biodiversity data managers who are unfamiliar with information science data practices of data standardization, the use of complex software to assist in the creation of standardized datasets can be a barrier to sharing data.

Since the ratification of the Darwin Core Standard (DwC) (Darwin Core Task Group 2009) by the Biodiversity Information Standards (TDWG) in 2009, many datasets have been published and shared through a variety of data portals. In the early stages of biodiversity data sharing, the protocol Distributed Generic Information Retrieval (DiGIR), progenitor of DwC, and later the protocols BioCASe and TDWG Access Protocol for Information Retrieval (TAPIR) (De Giovanni et al. 2010) were introduced for discovery, search and retrieval of distributed data, simplifying data exchange between information systems. Although these protocols are still in use, they are known to be inefficient for transferring large amounts of data (GBIF 2017). Because of that, in 2011 the Global Biodiversity Information Facility (GBIF) introduced the Darwin Core Archive (DwC-A), which allows more efficient data transfer, and has become the preferred format for publishing data in the GBIF network. DwC-A is a structured collection of text files, which makes use of the DwC terms to produce a single, self-contained dataset. Many tools for assisting data sharing using DwC-A have been introduced, such as the Integrated Publishing Toolkit (IPT) (Robertson et al. 2014), the Darwin Core Archive Assistant (GBIF 2010) and the Darwin Core Archive Validator. Despite promoting and facilitating data sharing, many users have difficulties using such tools, mainly because of the lack of training in information science in

© Salim J, Saraiva A. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the biodiversity curriculum (Convention on Biological Diversiity 2012, Enke et al. 2012). However, most users are very familiar with spreadsheets to store and organize their data, but the adoption of the available solutions requires data transformation and training in information science and more specifically, biodiversity informatics. For an example of how spreadsheets can simplify data sharing see Stoev et al. (2016).

In order to provide a more "familiar" approach to data sharing using DwC-A, we introduce a new tool as a Google Sheet Add-on. The Add-on, called *Darwin Core Archive Assistant Add-on* can be installed in the user's Google Account from the G Suite MarketPlace and used in conjunction with the Google Sheets application.

The Add-on assists the mapping of spreadsheet columns/fields to DwC terms (Fig. 1), similar to IPT, but with the advantage that it does not require the user to export the spreadsheet and import it into another software. Additionally, the Add-on facilitates the creation of a star schema in accordance with DwC-A, by the definition of a "**CORE_ID**" (e.g. occurrenceID, eventID, taxonID) field between sheets of a document (Fig. 2). The Add-on also provides an Ecological Metadata Language (EML) (Jones et al. 2019) editor (Fig. 3) with minimal fields to be filled in (i.e., mandatory fields required by IPT), and helps users to generate and share DwC-Archives stored in the user's Google Drive, which can be downloaded as a DwC-A or automatically uploaded to another public storage resource like a user's Zenodo Account (Fig. 4).

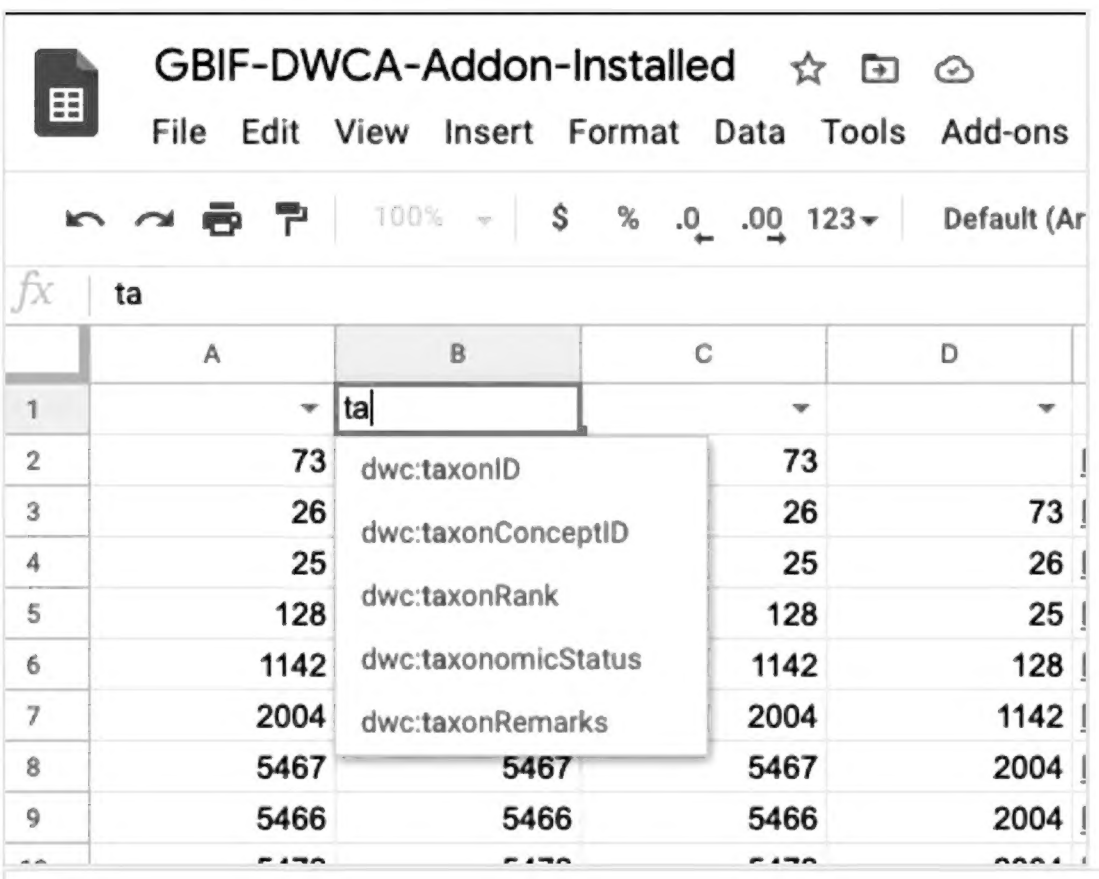

Figure 1.

An example of mapping sheet columns/fields to Darwin Core terms using the Darwin Core Archive Assistant Add-on.

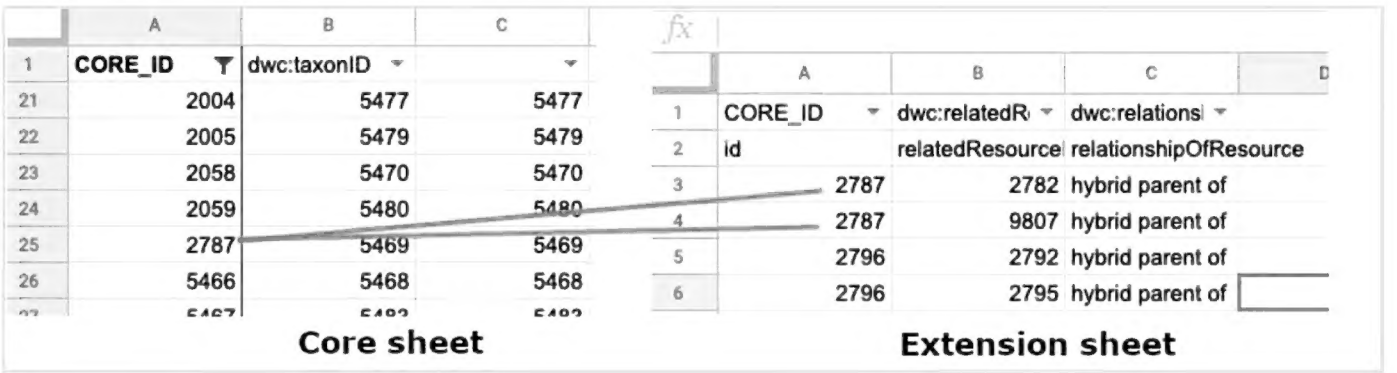

Figure 2.

An example of how to create a star schema using the Darwin Core Archive Assistant Add-on: the row in the left sheet (the **core sheet**) with CORE_ID 2787 is linked to the two rows of the right sheet (the **extension sheet**), creating a one-to-many relation between sheets, following the DwC-A nomenclature.

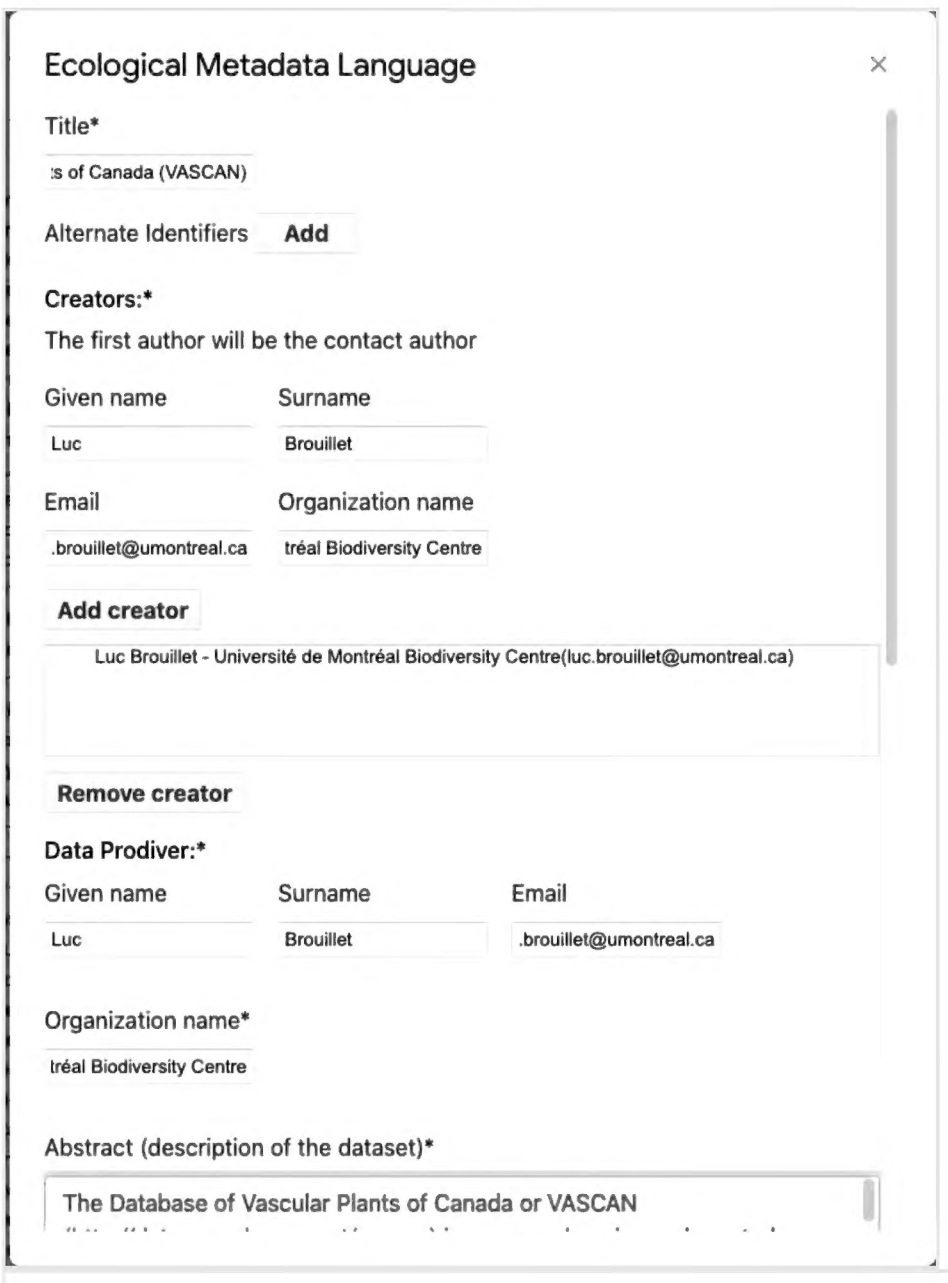

Figure 3.

The EML Editor of the Darwin Core Archive Assistant Add-on.

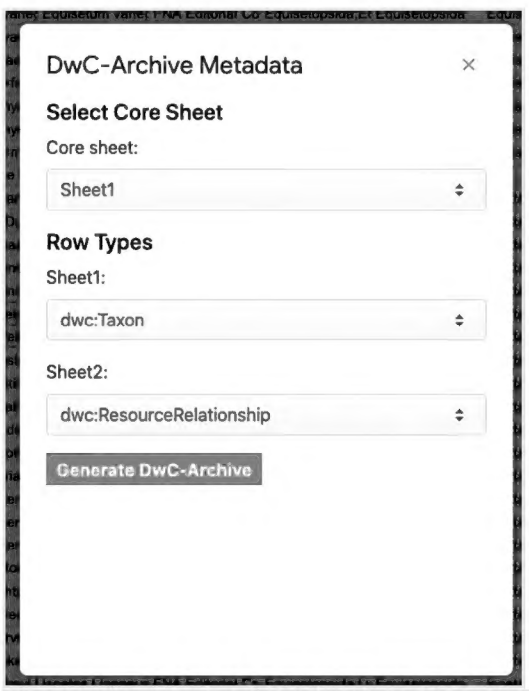

Figure 4.

Generating a Darwin Core Archive using the Darwin Core Archive Assistant Add-on: users have to select the "Core Sheet" and the Row Types (Darwin Core classes recognized by GBIF as cores) of each sheet (core and extensions).

We expect that the Google Sheet Add-on introduced here, in conjunction with IPT, will promote biodiversity data sharing in a standardized format, as it requires minimal training and simplifies the process of data sharing from the user's perspective, mainly for those users not familiar with IPT, but that historically have worked with spreadsheets. Although the DwC-A generated by the add-on still needs to be published using IPT, it does provide a simpler interface (i.e., spreadsheet) for mapping data sets to DwC than IPT. Even though the IPT includes many more features than the Darwin Core Assistant Add-on, we expect that the Add-on can be a "starting point" for users unfamiliar with biodiversity informatics before they move on to more advanced data publishing tools. On the other hand, Zenodo integration allows users to share and cite their standardized data sets without publishing them via IPT, which can be useful for users without access to an IPT installation. Additionally, we are working on new features and future releases will include the automatic generation of Global Unique Identifiers for shared records, the possibility of adding additional data standards and DwC extensions, integration with GBIF REST API and with IPT REST API.

Keywords

data sharing tool, Darwin Core, Darwin Core Archive, biodiversity informatics, spreadsheet

Presenting author

José Augusto Salim

Presented at

TDWG 2020

References

- Convention on Biological Diversiity (2012) A review of barriers to the sharing of biodiversity data and information, with recommendations for eliminating them. UNEP/CBD/COP/11/INF/8. URL: https://www.cbd.int/doc/meetings/cop/cop-11/information/cop-11-inf-08-en.pdf
- Darwin Core Task Group (2009) Darwin Core. Biodiversity Information Standards (TDWG). (Kampmeier G, review manager). URL: http://www.tdwg.org/standards/450
- De Giovanni R, Döring M, Güntsch A, Vieglais D, Hobern D, de la Torre J, Wieczorek J, Gales R, Hyam R, Blum S, Perry S (2010) TDWG Access Protocol for Information Retrieval (TAPIR), Version 1.0. Biodiversity Information Standards (TDWG). URL: http://www.tdwg.org/standards/449
- Enke N, Thessen A, Bach K, Bendix J, Seeger B, Gemeinholzer B (2012) The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —. Ecological Informatics 11: 25-33. https://doi.org/10.1016/j.ecoinf.2012.03.004
- GBIF (2010) Darwin Core Archive Assistant - Online tool. Copenhagen: Global Biodiversity Information Facility. URL: http://tools.gbif.org/dwca-assistant/
- GBIF (2017) Darwin Core Archives – How-to Guide, version 2.0, released on 9 May 2011, (contributed by Remsen D, Braak, K, Döring M, Robertson, T). https://github.com/gbif/ipt/wiki/DwCAHowToGuide. Accessed on: 2020-8-11.
- Jones M, OBrien M, Mecum B, Boettiger C, Schildhauer M, Maier M, Whiteaker T, Earl S, Chong S (2019) Ecological Metadata Language version 2.2.0. KNB Data Repository https://doi.org/10.5063/F11834T2
- Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P (2014) The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE 9 (8). https://doi.org/10.1371/journal.pone.0102623
- Stoev P, Smirnova L, Mergen P, Groom Q, De Wever A, Penev L, Pe'er I, Runnel V, Camacho A, Vincent T, Agosti D, Arvanitidis C, Bonet F, Saarenmaa H (2016) Data sharing tools adopted by the European Biodiversity Observation Network Project. Research Ideas and Outcomes 2 https://doi.org/10.3897/rio.2.e9390